Discussion of “The Shrinkage-Delinkage Trade-off”

Variational inference (VI) finds $q^* := \operatorname{argmin}_{q \in \mathcal{Q}} \mathrm{KL}\left(q || p\right)$ for an unknown target p.

What should $\mathcal{Q}$ be?

Discussion of "The Shrinkage-Delinkage Trade-off"

Variational inference (VI) finds $q^* := \mathrm{argmin}_{q \in \mathcal{Q}} \mathrm{KL}\left(q || p\right)$ for an unknown target p.

What should $\mathcal{Q}$ be?

Classical VI takes a simple $\mathcal{Q}$. Then $p \notin \mathcal{Q}$, but you get computational benefits!

Discussion of “The Shrinkage-Delinkage Trade-off”

Variational inference (VI) finds $q^* := \operatorname{argmin}_{q \in \mathcal{Q}} \mathrm{KL}\left(q || p\right)$ for an unknown target p.

What should $\mathcal{Q}$ be?

Classical VI takes a simple $\mathcal{Q}$. Then $p \notin \mathcal{Q}$, but you get computational benefits!

But when $p \notin \mathcal{Q}$, can get poor posterior approximations even in simple cases.

What to do?

Discussion of “The Shrinkage-Delinkage Trade-off”

Variational inference (VI) finds $q^* := \operatorname{argmin}_{q \in \mathcal{Q}} \mathrm{KL}\left(q || p\right)$ for an unknown target p.

What should $\mathcal{Q}$ be?

Classical VI takes a simple $\mathcal{Q}$. Then $p \notin \mathcal{Q}$, but you get computational benefits!

But when $p \notin \mathcal{Q}$, can get poor posterior approximations even in simple cases.

What to do?

1. Don’t care (“machine learning”)
    - Evaluate by other criteria than poterior approximations (e.g. prediction)
    - Maybe fine for some machine learning tasks

Discussion of "The Shrinkage-Delinkage Trade-off"

Variational inference (VI) finds $q^* := \operatorname{argmin}_{q \in \mathcal{Q}} \mathrm{KL}(q||p)$ for an unknown target p.

What should $\mathcal{Q}$ be?

Classical VI takes a simple $\mathcal{Q}$. Then $p \notin \mathcal{Q}$, but you get computational benefits!

But when $p \notin \mathcal{Q}$, can get poor posterior approximations even in simple cases.

What to do?

1. Don't care ("machine learning")
   - Evaluate by other criteria than poterior approximations (e.g. prediction)
   - Maybe fine for some machine learning tasks
2. Make $\mathcal{Q}$ more expressive ("modern VI")
   - Strong theoretical guarantees
   - High computational cost!

Discussion of “The Shrinkage-Delinkage Trade-off”

Variational inference (VI) finds $q^* := \mathrm{argmin}_{q \in \mathcal{Q}} \mathrm{KL}\left(q||p\right)$ for an unknown target p.

What should $\mathcal{Q}$ be?

Classical VI takes a simple $\mathcal{Q}$. Then $p \notin \mathcal{Q}$, but you get computational benefits!

But when $p \notin \mathcal{Q}$, can get poor posterior approximations even in simple cases.

What to do?

1. Don’t care (“machine learning”)
    - Evaluate by other criteria than poterior approximations (e.g. prediction)
    - Maybe fine for some machine learning tasks
2. Make $\mathcal{Q}$ more expressive (“modern VI”)
    - Strong theoretical guarantees
    - High computational cost!
3. Try to capture important properties of p with simple $\mathcal{Q}$
    - Begins with understanding how things go wrong (**this paper**!)
    - Hope to have our cake and eat it too (e.g. marginals *and* easy computation)
    - Much harder! But important, with big potential benefits

I would love to see more work like this!

Discussion of “The Shrinkage-Delinkage Trade-off”

Restricted variational families (mean field) can lead to poor posterior approximations.

Two very common approaches to VI are:

- “Machine learning”: Ignore it (evaluate using some other criteria, like prediction)
- “Modern VI”: Use more expressive families (at a computational cost)

This paper tries to understand *how* the restricted family goes wrong. *fire emoji*

Discussion of “The Shrinkage-Delinkage Trade-off”

Restricted variational families (mean field) can lead to poor posterior approximations.

Two very common approaches to VI are:

- “Machine learning”: Ignore it (evaluate using some other criteria, like prediction)
- “Modern VI”: Use more expressive families (at a computational cost)

This paper tries to understand *how* the restricted family goes wrong. *fire emoji*

This paper’s most (initially) surprising conclusion is probably this:

Theorem 3.6

Let the target distribution has the constant ε-correlation matrix.
As the dimension n of the matrix goes to infinity:

- Each marginal mean field variance is wrong by ε
- The per-component entropy gap $\to 0$

Discussion of “The Shrinkage-Delinkage Trade-off”

Restricted variational families (mean field) can lead to poor posterior approximations.

Two very common approaches to VI are:

- “Machine learning”: Ignore it (evaluate using some other criteria, like prediction)
- “Modern VI”: Use more expressive families (at a computational cost)

This paper tries to understand *how* the restricted family goes wrong. *fire emoji*

This paper’s most (initially) surprising conclusion is probably this:

Theorem 3.6

Let the target distribution has the constant ε-correlation matrix.
As the dimension n of the matrix goes to infinity:

- Each marginal mean field variance is wrong by ε
- The per-component entropy gap $\to 0$

The key is “per-component entropy gap” means **entropy difference / n**.

In fact, one can show $\text{Entropy gap} = O(\log n) \to \infty$. **Why is n the “right” scaling?**

Discussion of “The Shrinkage-Delinkage Trade-off”

Restricted variational families (mean field) can lead to poor posterior approximations.

Two very common approaches to VI are:

- “Machine learning”: Ignore it (evaluate using some other criteria, like prediction)
- “Modern VI”: Use more expressive families (at a computational cost)

This paper tries to understand *how* the restricted family goes wrong. *fire emoji*

This paper’s most (initially) surprising conclusion is probably this:

Theorem 3.6

Let the target distribution has the constant ε-correlation matrix.
As the dimension n of the matrix goes to infinity:

- Each marginal mean field variance is wrong by ε
- The per-component entropy gap $\to 0$

The key is “per-component entropy gap” means **entropy difference / n**.

In fact, one can show $\text{Entropy gap } = O(\log n) \to \infty$. **Why is n the “right” scaling?**

Why do the relative values across dimensions of the entropy gap matter?

It’s clear why variance is useful. Less so the entropy gap, especially as n changes.